IOWA STATE UNIVERSITY
Digital Repository

CSAFE Publications

Center for Statistics and Applications in Forensic Evidence

4-2019

What does a match mean? A Framework for Understanding Forensic Comparisons

Robin Mejia
Carnegie Mellon University

Maria Cuellar
University of Pennsylvania

Dana M. Delger
The Innocence Project

Bill Eddy
Carnegie Mellon University

Follow this and additional works at: https://lib.dr.iastate.edu/csafe_pubs

 Part of the Legal Studies Commons

Recommended Citation

Mejia, Robin; Cuellar, Maria; Delger, Dana M.; and Eddy, Bill, "What does a match mean? A Framework for Understanding Forensic Comparisons" (2019). *CSAFE Publications*. 66.
https://lib.dr.iastate.edu/csafe_pubs/66

This Article is brought to you for free and open access by the Center for Statistics and Applications in Forensic Evidence at Iowa State University Digital Repository. It has been accepted for inclusion in CSAFE Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

What does a match mean? A Framework for Understanding Forensic Comparisons

Abstract

On 11 March 2004, terrorists in Madrid, Spain detonated bombs on several commuter trains. In total, 191 people were killed and 1,400 were injured. After the bombing, examiners from the Federal Bureau of Investigation (FBI) identified a latent fingerprint found on a bag containing detonators and explosives as coming from an Oregon lawyer named Brandon Mayfield. Mayfield was arrested and held as a material witness for two weeks, until the Spanish National Police determined that the print did not, in fact, come from Mayfield, but from another man living in Spain.

Disciplines

Legal Studies

Comments

The following article is published as Mejia, Robin, Maria Cuellar, Dana Delger, and Bill Eddy. "What does a match mean? A framework for understanding forensic comparisons." *Significance* 16, no. 2 (2019): 25-28. Posted with permission of CSAFE.

This article is available at Iowa State University Digital Repository: https://lib.dr.iastate.edu/csafe_pubs/66

What does a match mean? A framework for understanding forensic comparisons

By **Robin Mejia**, **Maria Cuellar**, **Dana Delger** and **Bill Eddy**

On 11 March 2004, terrorists in Madrid, Spain detonated bombs on several commuter trains. In total, 191 people were killed and 1,400 were injured. After the bombing, examiners from the Federal Bureau of Investigation (FBI) identified a latent fingerprint found on a bag containing detonators and explosives as coming from an Oregon lawyer named Brandon Mayfield. Mayfield was arrested and held as a material witness for two weeks, until the Spanish National Police determined that the print did not, in fact, come from Mayfield, but from another man living in Spain.

How did this happen? A "senior fingerprint examiner" at the FBI, who made the original identification, "'consider[ed] it to be a 100% identification' of Mayfield". The match was verified by the unit chief of the FBI's Latent Print Unit, "a retired FBI fingerprint examiner with over thirty years of experience", and an independent fingerprint examiner "widely considered a leader in the profession".[1] After the error was uncovered, the Office of the Inspector General for the United States Department of Justice investigated Mayfield's case. Among other findings, it concluded that "the unusual similarity of details on the fingers of Mayfield and the true source of the print ... confused the FBI Laboratory examiners, and was an important factor contributing to the erroneous identification" (bit.ly/2Ezvbwr).

Mayfield is far from the only person to suffer from a miscarriage of justice. Since 1989, more than 2,000 individuals have been exonerated after having been wrongfully convicted, according to the National Registry of Exonerations (bit.ly/2EzTlXz). Disturbingly, around a quarter of those cases included "false or misleading forensic evidence".

One of the issues that can lead to errors in forensic analysis (as is apparent in Mayfield's case) is the way in which examiners deal with uncertainty. In 2016, a report from the President's Council of Advisors on Science and Technology (PCAST) noted that forensic examiners frequently state that their conclusions about forensic evaluations are "100 percent certain"; have error rates that are "essentially zero", "vanishingly small", or "microscopic"; or have a chance of error so remote as to be a "practical impossibility" (bit.ly/2EFU89o).

To a statistician, these characterisations of error in a process of human matching sound vague and implausible, but to a jury member or a judge they can sound very convincing. This is especially true when such characterisations come from an expert witness. What is not clear from the confidence statements is that they often reflect only the forensic analyst's opinion about whether two items match or not, and they fail to take into account the value of that match. As demonstrated by Mayfield's case (and many others), similarity alone is not sufficient to understand the value of an item of evidence. The lack of a proper foundation to discuss uncertainty in those values in forensic conclusions is likely to have led to wrongful convictions.

© 2019 The Royal Statistical Society

A two-step process

To assess the probative value of a piece of evidence – how important it is in connecting a suspect to a crime – we need to understand two things: whether the crime scene and suspect-associated evidence appear to be similar, and, if so, what that similarity means. More specifically, we need to know (1) whether the evidence from the crime scene "matches" a sample from a suspect by some defined match criterion, and (2) how probable it is that a match by that criterion would occur by chance. Each of these components can be framed as a statistical question with an answer that can be estimated from data.

It is important to differentiate between steps 1 and 2. Saying that two hairs are visually indistinguishable is different than saying that they are visually indistinguishable and therefore the one from the crime scene must have come from the suspect's head. When analysts opine that two specimens come from the same source with near or complete certainty, they are frequently conflating steps 1 and 2, assuming that only that source could have produced the measured characteristics.

But let us return to step 1. Today, how the level of similarity between items is established varies by evidence type. Ideally, some predefined set of characteristics should be measured on each piece of evidence, and those characteristics compared to each other. In some cases, this step is fairly objective and automated. For example, the similarity may be determined by a chemical analysis, using mass spectrometry to determine the trace element profile of a bullet or glass shard. In other cases, markings may initially be compared by an automated computer algorithm, as is the case with fingerprints or some ballistics comparisons (see page 31). In many cases, however, a final determination of whether two items match is made by subjective human judgement of trained examiners.

In some disciplines the entire technique is subjective: comparisons of hair or bite marks, for example. In these cases, the characteristics and measurements may not even be defined in advance of the analysis. Such a procedure can contribute to confirmation bias; once a conclusion is tentatively reached, there is a tendency for individuals to focus on evidence that supports the existing belief. (This phenomenon is observed across fields, and is one of the reasons why predefined analysis protocols are required for many kinds of studies.) The 2016 PCAST report emphasised the importance of standardising the measurement and comparison of samples and developing automated methods wherever possible. In addition, rather than relying on a binary "match"–"no match" decision, one could also consider developing a similarity score, or level of *matchingness*, for this step.

Robin Mejia holds a special faculty appointment in the Department of Statistics and Data Science at Carnegie Mellon University, where she works on quantitative assessments of human rights issues and improving the validity and reliability of forensic science procedures.

Maria Cuellar is an assistant professor in the University of Pennsylvania Criminology Department. She received her PhD in the joint statistics and public policy programme at Carnegie Mellon University, and she later completed a postdoctoral fellowship at Penn.

Step 2, determining what a match or given level of similarity means, is arguably the more challenging step of the process. To estimate how likely it is that two samples could match at a given level by chance, one needs to know how common are the characteristics used to determine the match. To estimate that, one needs data on the population from which the samples came. For example, in the case of single-source or simple-mixture DNA samples, one needs to know the prevalence of different DNA profiles in a population. Fortunately, DNA researchers have collected this kind of data in data sets that are probability samples from the populations of interest (see, for example, strbase.nist.gov). These data allow examiners to compute things like random match probabilities, to quantify the probability of chance matches.

If one wanted to reach a similarly supportable conclusion about the chance of seeing specific markings on a cartridge case, one would need to know about the population of guns in use in, for example, the country or the region where the crime occurred. If one wanted to know how likely it was that a glass shard came from a random car headlight, one would need to know about the population of glass headlights in the country or perhaps a certain county. Unfortunately, population reference data are not available in these domains.

Of course, the question remains: how much does this really matter in practice? Without the data, it is hard to know. However, we can at least consider what we do not know.

Imagine some bullets are found at a crime scene, and a suspect also has bullets at her apartment. An analyst could analyse the elemental profile of the bullets. Then, the analyst could compare the measurements of those trace elements and determine whether those measurements match, within some level of error. If the crime scene and suspect bullets are clearly dissimilar, that alone might be enough to rule out the possibility that the two items are related. However, if they are similar, more information is needed to assess how important that finding is. A match would mean very different things if there are 100,000 similar bullets in a city with the same trace element profile than if there are only 20 bullets in the world with the same elemental signature. But a jury which heard only that the examiner was 100% certain that the bullets were a match would have no way of telling the difference.

It is obviously impossible to examine every bullet in the world, or even in a city. However, it is possible to estimate population rates for characteristics from sampled data. If, for example, we had a database of bullets that was representative

> Saying that two hairs are visually indistinguishable is different than saying that they are visually indistinguishable and therefore the one from the crime scene must have come from the suspect's head

of all bullets that existed in the region (or country) where the crime occurred, this could help us understand the probability of a chance match. Unfortunately, for most areas of forensic science where such comparisons are made, data to estimate this probability are lacking. This includes comparative bullet lead analysis, the technique just described (see page 13 for more on this).

In most domains, there are no population reference data sets. Even in domains where large databases exist – for example, as in fingerprints and ballistics – the data sets are not probability samples from an underlying population of interest and are not substitutes for them. That said, understanding the frequency of evidential characteristics in a very large but not representative data set could still provide important information. If, for example, we knew that certain types of fingerprint characteristics were incredibly common in the FBI's Next Generation Identification (NGI) system, which stores fingerprint and other biometric information, we would know that the possibility of a random match on those characteristics is not small. However, criminal databases such as the NGI and the National Integrated Ballistics Information Network are not available to researchers. Currently, only in the case of DNA do we have the kind of data we need to estimate the probability of a random match.

'Black box' studies are not enough

In addition to calling for more objective methods, the 2016 PCAST report discussed ways to improve our understanding of subjective methods in the interim, such as through "black box" studies to generate error rates (previously discussed on page 23). These studies invite examiners to "analyze samples and render opinions about the origin or similarity of samples". They can be valuable in assessing how well a given technique works in practice, but their "black box" nature does not allow us to distinguish between human error in matching samples and very similar samples that happen to be from different sources. Therefore, they address a different question than the one we are discussing here.

To estimate an error rate, one presents an examiner with a series of evidential comparisons where the truth of whether a crime scene sample is related to a subject is known. One then measures how many times the examiner gets the answer wrong, and divides that by the total number of comparisons. This is a straightforward measure to compute, but it does not fully tease out either examiner skill or the chance of a coincidental match. Rather, whether an examiner gets an answer right depends on several factors: the examiner's skill, the inherent variation in a class of evidence, and the particular set of evidence they are asked to compare.

If one wishes to truly understand the probative value of a particular piece of evidence, it is important to understand the probability that two items will match by chance under the criteria an examiner uses. In other words, while black box studies can tell us something important about the likelihood that an examiner is correct or incorrect in his or her conclusion that a piece of trace evidence matches a sample from a suspect, they cannot tell us everything. Data on the frequency of characteristics in a population are needed to fully inform the jury or other fact-finder about the value of the particular piece of evidence they are considering – and that value is, ultimately, what they are being asked to consider.

Dana M. Delger is a staff attorney in the Strategic Litigation Unit of the Innocence Project, which uses the courts strategically to address the leading causes of wrongful conviction, including eyewitness misidentification and the misapplication of forensic sciences.

Bill Eddy is the John C. Warner Professor of Statistics, Emeritus, at Carnegie Mellon University, with appointments in the Department of Biological Sciences, Machine Learning Department, and the Center for the Neural Basis of Cognition.

For most areas of forensic science where such comparisons are made, data to estimate the probability of a chance match are lacking

Transparent analyses

Essentially, what we (and participants in the justice system) want to understand is the probability of seeing the evidence found at a crime scene if the suspect is innocent: P(Evidence | Innocent). A strong argument can be made that the question that a juror is most likely to care about is the probability that the suspect is guilty given that a piece of evidence is observed at a crime scene: P(Guilty | Evidence). This is much more difficult to estimate, so we focus on what we can estimate from the evidence and knowledge of the population from which that evidence comes, which is the probability of seeing the evidence given the suspect's innocence. This is the probability of seeing the characteristics observed in the suspect evidence and in a random sample of that type of material.

In some areas, there has been a move to use likelihood ratios to summarise the probative value of evidence (see page 14). Using Bayes' rule, we can write that as

$$\frac{\text{P(Guilty} \mid \text{Evidence)}}{\text{P(Innocent} \mid \text{Evidence)}} = \frac{\text{P(Evidence} \mid \text{Guilty)}}{\text{P(Evidence} \mid \text{Innocent)}} \times \frac{\text{P(Guilty)}}{\text{P(Innocent)}}.$$

In other words, posterior odds = likelihood ratio × prior odds. In theory, the prior odds are known by the juror or judge, and the likelihood ratio is estimated by the expert witness. The trier of fact then combines these two using Bayes' rule (intuitively) to obtain a posterior odds, which determines how much more likely the suspect is to be guilty than innocent, given the evidence. Thus, the trier of fact has updated his or her belief that the individual is guilty after learning the likelihood ratio from the expert.

Although the ratio form of Bayes' rule provides a transparent way to interpret how individuals update their beliefs, there is an ongoing debate about the proper way to estimate the likelihood ratio. One of the concerns is that using different databases to set the denominator, P(Evidence | Innocent), could yield drastically different estimates of the ratio. Again, ideally the denominator should be estimated from population reference data.

Our point is not that everything can and should be like DNA, but that we need to understand the value each type of evidence provides

Furthermore, the selection of probabilistic models for the likelihood ratio – a part of the process that is up to a statistician or expert to select – also affects the results. So, even if estimating the likelihood ratio is the best means of communicating this information in the abstract, whoever does this must tread carefully in, and be very clear about, the analysis selected and its assumptions, and those choices must be clearly communicated to the fact-finder. It also bears repeating that a move to likelihood ratios does not *on its own* solve the problem this article addresses: the lack of data for calculating the chance of a random match. Likelihood ratios cannot absolve the examiner of needing to have empirical data supporting his or her claim about the value of a match.

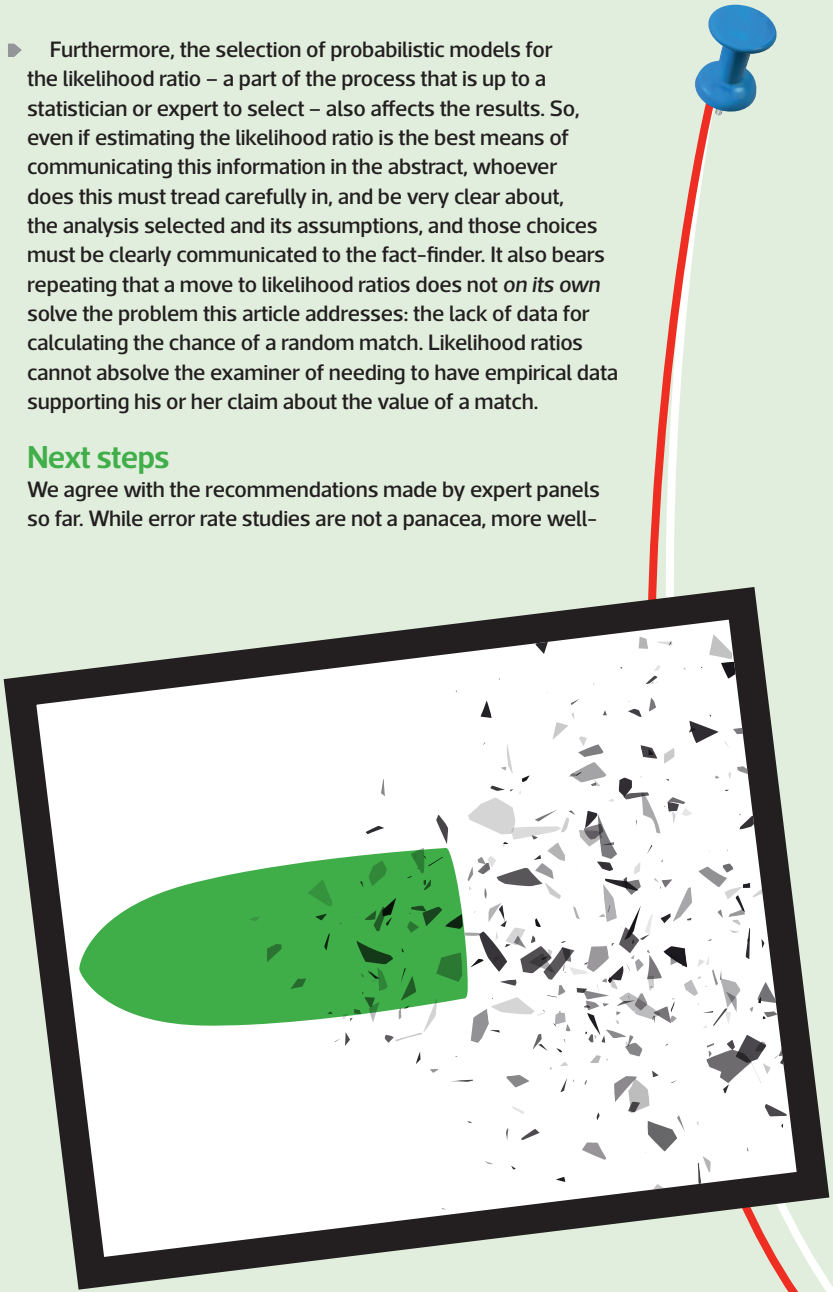
Next steps

We agree with the recommendations made by expert panels so far. While error rate studies are not a panacea, more well-designed studies would greatly aid our understanding of how often errors are occurring in practice. Such studies still add important information to our understanding of the value of forensic science evidence. For example, it is now widely accepted that errors in DNA analysis are overwhelmingly due to examiner error rather than chance matches. Most likely, population reference studies in other domains will find different rates of chance matches. Our point is not that everything can and should be like DNA, but that we need to understand the value each type of evidence provides.

For now, it is clear that there are many obstacles to properly expressing the uncertainty in a forensic match. First, forensic analysts and scientists promoting the reform of forensic science must understand that two quantities are necessary: the probability that the evidence from the crime scene matches a sample from a suspect *and* the probability that this match could occur by chance. Second, to estimate these probabilities, adequate databases of forensic evidence must be generated. Finally, experts must present their analyses of the data in a transparent way, by describing their assumptions for modelling and data selection.

We are encouraged by efforts to build forensic databases that enable initial investigations into these questions, including the National Institute of Standards and Technology (NIST) Ballistic Toolmark Research Database (tsapps.nist.gov/NRBTD), and the forthcoming forensic fingerprint database under development at NIST. Other researchers, including our colleagues in the Center for Statistics and Applications in Forensic Evidence, are developing databases of steganography images and elemental glass composition, and some of us are planning a database of tool-mark striations. All of these efforts will help us begin to assess the characteristics that provide the most information about different types of evidence. However, the "begin" in the preceding sentence is important. A huge need remains for access to larger data sets, such as those maintained by law enforcement, and for intentional collection of sampled data designed to represent a population.

Until we have access to such data sets, it is impossible to know how many cases like Brandon Mayfield's are out there but not discovered, or how many techniques suffer from the issues that plagued bullet lead analysis (see page 13). Conversely, it is impossible to fully characterise the value that many analyses provide. We fully expect that many forensic comparison techniques provide a great deal of useful information, and access to population reference data would enable us to characterise their value. Addressing these challenges will enable forensic analysts to present the uncertainty in their conclusions, better equipping judges and juries to use that evidence in their determination of whether someone is innocent or guilty of a crime. ■



Reference

1. Cole, S. A. (2005) More than zero: Accounting for error in latent fingerprint identification. *Journal of Criminal Law and Criminology*, **95**, 985–1078.